

Decomposition of bipartite and multipartite unitary gates into the product of controlled unitary gates

Lin Chen^{1,2} and Li Yu^{2,3,*}

¹*Department of Mathematics, Beijing University of Aeronautics and Astronautics, Beijing 100191, P. R. China*

²*Singapore University of Technology and Design, 20 Dover Drive, Singapore 138682*

³*National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan*

(Dated: March 19, 2015)

We show that any unitary operator on the $d_A \times d_B$ system ($d_A \geq 2$) can be decomposed into the product of at most $4d_A - 5$ controlled unitary operators. The number can be reduced to $2d_A - 1$ when d_A is a power of two. We also prove that three controlled unitaries can implement a bipartite complex permutation operator, and discuss the connection to an analogous result on classical reversible circuits. We further show that any n -partite unitary on the space $\mathbb{C}^{d_1} \otimes \cdots \otimes \mathbb{C}^{d_n}$ is the product of at most $2 \prod_{j=1}^{n-1} (2d_j - 2) - 1$ controlled unitary gates, each of which is controlled from $n - 1$ systems. We also decompose any bipartite unitary into the product of a simple type of bipartite gates and some local unitaries. We derive dimension-independent upper bounds for the CNOT-gate cost or entanglement cost of bipartite permutation unitaries (with the help of ancillas of fixed size) as functions of the Schmidt rank of the unitary. It is shown that such costs under a simple protocol are related to the log-rank conjecture in communication complexity theory via the link of nonnegative rank.

PACS numbers: 03.65.Ud, 03.67.Lx, 03.67.Mn

I. INTRODUCTION

The implementation of unitary operations is a key task in quantum information processing. Unitary operators can be implemented by passive linear optical devices [1]. It is known that any unitary operation on two or more parties can be decomposed into the product of controlled unitary gates [2, 3]. Two-qubit controlled unitaries can be implemented with high coherence and dynamical coupling [4]. Suppose that a bipartite unitary U on systems A, B is the product of k bipartite controlled unitaries, interspersed with local unitaries [5]. We call the integer k as the *bipartite depth* of the circuit under the bipartite cut $A-B$. The depth, width and total number of basic gates are often quantities of interest in quantum circuit design, where the basic gates refer to some fixed type of two-qubit gates such as the controlled-NOT (CNOT) gate. For implementing the same unitary operation, it is conceivable that there may be a tradeoff between the depth and the total number of basic gates. Nonetheless the bipartite depth does give an upper bound for the total number of basic gates, as discussed in Sec. V of this paper. The nonlocal gates need much longer time than local gates to implement, because the systems may be far from each other. Then the bipartite depth is a rough measure of time needed by the circuit. By allowing local unitary freedom in the definition of controlled unitaries (in Sec. II), from now on we will drop the phrase “interspersed with local unitaries” from the definition of the bipartite depth.

We define the *bipartite depth* of a given bipartite uni-

tary U as the minimum bipartite depth among all unitary circuits for U that do not use ancillas. Formally, it is

$$c(U) := \min\{k | U = U_1 U_2 \cdots U_k, U_i \in \mathcal{S}\}, \quad (1)$$

where \mathcal{S} is the set of bipartite controlled unitaries on the same space that U acts on. Studying the bounds for $c(U)$ and the corresponding decomposition of U is the main problem in this paper. Indeed, it is a special case of the problem of quantum circuit decomposition using *general* controlled unitaries with the help of local unitaries. It is special in the sense that there are only two systems but the general problem allows many systems. There has been study on decompositions using CNOT or other two-qubit controlled gates, or specific classes of two-qubit controlled gates [2, 3, 6, 7]. For example, Shende *et al.* [8] shows that any three-qubit unitary can be written as the product of 20 CNOT gates and some one-qubit unitaries. Another motivation to study the problem is to better understand the structure of nonlocal unitaries and the resources needed to implement them, see the comment just before Section III A.

We restrict to bipartite controlled gates as the type of nonlocal gates in the definition of bipartite depth for the following reasons. First, it is easy to define, and a smaller class of gates seems not powerful enough. It is hard to find a larger class of easily definable gates that do not include all bipartite gates. The Fourier hierarchy [9] concerns the number of tensor products of Hadamard gates in a circuit that also contains basis-preserving gates. The basis-preserving gates are also called the complex permutation gates, and are discussed later in this paper. They permute among computational-basis states and apply a phase to each state. However the basis-preserving gates are generally nonlocal with respect to a bipartite partition of the qubits. If we modify the definition of Fourier

*Electronic address: liyu@nii.ac.jp

hierarchy and apply it to the bipartite scenario so as to allow some finite set of bipartite gates and arbitrary local gates, then such a set of bipartite gates would have a discrete set of entangling power, which is not desirable for defining a smooth depth measure. Second, the controlled unitaries are analogous to some components in protocols with local operations and classical communication (LOCC). They are a major type of protocols studied in quantum information theory. The LOCC protocols often allow projective measurements on some subsystems. A projective measurement and the subsequent classically controlled unitary operations can be made part of a coherent quantum circuit by rewriting them as a controlled unitary. Thus our measure is analogous to the rounds of classical communications in such protocols.

Generally we consider unitaries acting on $d_A \times d_B$ dimensional systems. The results of [2, 3] imply that $c(U) \leq \mu d_A^4$ when $d_A = d_B$, where μ is a positive constant, and the type of bipartite controlled gates used are limited to controlled-increment gates. In Theorem 4, we obtain a tighter bound $c(U) \leq 4d_A - 5$ for arbitrary d_A, d_B at the cost of allowing the use of arbitrary controlled-unitary gates in the decomposition. The same theorem shows that the bound can be further reduced to $2d_A - 1$ when d_A is a power of 2. We also prove that $c(U) \leq 3$ when U is a complex permutation matrix in Theorem 7, based on the concept of absolute singularity studied in Lemma 6. This result is applied to classical reversible circuits [10, 11] in Corollary 8. The above results are based on the sandwich form of bipartite unitaries, constructed in Definition 2 and Lemma 3. We further generalize our observation to multipartite systems based on the generalized sandwich form. We show that any n -partite unitary on the space $\mathbb{C}^{d_1} \otimes \cdots \otimes \mathbb{C}^{d_n}$ has a generalized $[2 \prod_{j=1}^{n-1} (2d_j - 2) - 1]$ -sandwich form in Proposition 9. We also propose a more efficient generalized sandwich form for $n = 4$ in Proposition 10. In Proposition 11, we show that any n -partite complex permutation unitary has a generalized $(2^n - 1)$ -sandwich form composed of controlled-complex-permutation unitaries.

We also discuss the decomposition of any unitary gate using “standard” gates proposed in Definition 12. They effectively only act on two qubits as controlled unitaries, and may be more easily carried out in experiments. We show that any bipartite unitary is the product of $2(d_A - 1)^2 \lfloor \frac{d_B}{2} \rfloor + (2d_A - 3)(d_B - 1) \lfloor \frac{d_A}{2} \rfloor$ standard gates interspersed with local unitaries in Proposition 15. The number reduces to three for $d_A = d_B = 2$, which is the smallest number of controlled unitaries needed for the decomposition of two-qubit unitary gates [12]. In Sec. VI we discuss the relationship between the Schmidt rank of the unitary and the number of controlled unitaries needed to decompose it. We give a class of examples where the number of controlled unitaries is upper bounded by a constant, but the Schmidt rank of the target unitary is arbitrarily large.

The rest of the paper is organized as follows. In Sec. II we introduce some definitions and preliminary knowl-

edge. In Sec. III we study the decomposition of bipartite unitary operators using controlled unitaries, and comment on the connections with results in the literature. In Sec. IV we define the “controlled-type” multipartite unitaries and discuss the decomposition of multipartite operators into the product of these gates. We also show that three controlled-permutation matrices are enough to decompose any complex permutation matrix. In Sec. V we define the standard gates and discuss the decomposition of bipartite unitaries using these gates and local unitaries. In Sec. VI we discuss the relationship between the Schmidt rank of the unitary and the form of the decomposition, and we discuss bipartite permutation unitaries in particular. In Sec. VII we discuss the use of local ancillas. We conclude in Sec. VIII.

II. PRELIMINARIES

In this section we introduce the preliminary knowledge used in the paper. Denote the computational-basis states of the bipartite Hilbert space $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$ by $|i, j\rangle, i = 1, \dots, d_A, j = 1, \dots, d_B$. Let I_A and I_B be the identity operators on the spaces \mathcal{H}_A and \mathcal{H}_B , respectively. Any bipartite unitary gate U acting on \mathcal{H} has *Schmidt rank* (denoted as $\text{Sch}(U)$) equal to n if there is an expansion of the form $U = \sum_{j=1}^n A_j \otimes B_j$ where the $d_A \times d_A$ matrices A_1, \dots, A_n are linearly independent, and the $d_B \times d_B$ matrices B_1, \dots, B_n are also linearly independent. An equivalent definition is in [13, 14], where it is called the operator-Schmidt rank. Next, U is a *controlled unitary gate*, if U is equivalent to $\sum_{j=1}^{d_A} |j\rangle\langle j| \otimes U_j$ or $\sum_{j=1}^{d_B} V_j \otimes |j\rangle\langle j|$ via local unitaries. To be specific, U is a controlled unitary from the A or B side, respectively. In particular, U is controlled in the computational basis from the A side if $U = \sum_{j=1}^{d_A} |j\rangle\langle j| \otimes U_j$. Bipartite unitary gates of Schmidt rank two or three are in fact controlled unitaries [15–17]. We have generalized controlled unitaries to block-controlled unitary gates [16]. We split the space \mathcal{H}_A into a direct sum: $\mathcal{H}_A = \oplus_{i=1}^m \mathcal{H}_i$, $m > 1$, $\text{Dim } \mathcal{H}_i = m_i$, and $\mathcal{H}_i \perp \mathcal{H}_j$ for distinct $i, j = 1, \dots, m$. Then U is a *block-controlled unitary (BCU) gate controlled from the A side*, if U is locally equivalent to $\sum_{i=1}^m \sum_{j,k=1}^{m_i} |u_{ij}\rangle\langle u_{ik}| \otimes V_{ijk}$ where $\{|u_{i,1}\rangle, \dots, |u_{i,m_i}\rangle\}$ is an orthonormal basis of \mathcal{H}_i . Note that the V_{ijk} are not necessarily unitary. By definition every controlled unitary with $d_A, d_B \geq 2$ is a BCU. The BCU will be used in the proof of Theorem 4, as well as in the decomposition of any bipartite unitary into the product of three BCUs in Corollary 5.

III. DECOMPOSITION OF BIPARTITE UNITARY OPERATORS

It is known [12] that three controlled gates are sufficient and necessary for the decomposition of a general

two-qubit unitary, and there is always a decomposition using 3 CNOT gates and some one-qubit unitaries. For implementing a two-qubit SWAP gate by local unitaries and some number of CNOT gates without the use of ancillas (this condition of no ancillas is implied throughout the paper unless stated otherwise), three CNOT gates are necessary and sufficient [12]. We generalize this fact to the SWAP gates of arbitrary dimension.

Lemma 1 *Denote the two-qudit SWAP gate acting on $d \times d$ system as SWAP_d . Then*

- (i) *the product of the SWAP_d gate and any controlled unitary has Schmidt rank d^2 ;*
- (ii) *For implementing a SWAP_d gate by local unitaries and some number of controlled unitary gates, three controlled unitaries are necessary and sufficient.*

Proof. (i) There are orthonormal bases of \mathcal{H}_A and \mathcal{H}_B (denoted by $\{|i\rangle\}_A$ and $\{|j\rangle\}_B$) such that the matrix representation of the SWAP_d gate in such bases has elements of the form $\langle i|_A \langle j|_B U|k\rangle_A |l\rangle_B = \delta_{il} \delta_{jk}$. Because the SWAP_d gate effectively performs the physical swap of two systems, which is basis-independent, the above particular matrix representation is invariant under simultaneous unitary similarity transform (simultaneous unitary change of basis) on the two local systems. Then assertion (i) follows from straightforward computation, by writing the matrix for the SWAP_d gate in the form above and assuming one of the local bases is the local controlling basis for the controlled unitary.

(ii) Any controlled unitary on \mathcal{H} has Schmidt rank at most d . It follows from assertion (i) that the SWAP_d gate is the product of at least three controlled unitaries. It is known that the SWAP_d gate is the product of three controlled unitary gates [18]. So assertion (ii) holds. This completes the proof. \square

For the two-qubit SWAP gate, using the general controlled unitaries in its decomposition does not save any controlled unitary compared to using CNOT gates. One might expect that this is the general case, i.e., the implementation of a bipartite unitary is the same when we use controlled unitaries or only CNOT gates. However, the two-qubit gate $\exp(i a \sigma_1 \otimes \sigma_1)$ with the Pauli matrix $\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ any $a \neq k\pi/4$, $k \in \mathbb{Z}$ cannot be implemented using one CNOT gate and single qubit gates only, since the entangling power of such gate is not equal to that of the CNOT gate. We will show in Theorem 4 that for the general $d \times d$ bipartite system that using controlled unitaries might be better than the d -dimensional CNOT gates, in the sense that they require fewer such two-qudit gates. For this purpose we introduce a special decomposition of bipartite unitaries.

Definition 2 (i) *We refer to the m -sandwich form of a bipartite unitary U , in the sense that $U = U_1 U_2 \cdots U_m$, where each U_i is a controlled unitary, being controlled in the computational basis on the respective Hilbert space,*

and the party that does the controlling alternates between A for odd i and B for even i .

(ii) *We refer to the m -A form of a bipartite unitary U , in the sense that $U = U_1 U_2 \cdots U_m$, where any U_i is a controlled unitary controlled from the A side.*

Using this definition we present the following result as the first step to our question.

Lemma 3 (i) *Any $2 \times d_B$ unitary has a 3-sandwich form;*
(ii) *Any $2 \times d_B$ unitary has a 3-A form;*
(iii) *There exists a 2×2 unitary that cannot be the product of two controlled unitaries.*

Proof. (i) For any $2 \times d_B$ unitary M , there are two local unitaries E, F on \mathcal{H}_B such that $M = (I_A \otimes E)U(I_A \otimes F)$, where $U = \sum_{i,j=0}^1 |i\rangle\langle j| \otimes U_{ij}$ and U_{00} is a $d_B \times d_B$ diagonal matrix. Since U is unitary, the columns of U_{10} are pairwise orthogonal, and the rows of U_{01} are also pairwise orthogonal. Let V, W be two $d_B \times d_B$ unitaries such that both VU_{10} and $U_{01}W$ are diagonal matrices with all elements real and non-negative. Let $U_1 = |0\rangle\langle 0| \otimes I_B + |1\rangle\langle 1| \otimes V$ and $U_2 = |0\rangle\langle 0| \otimes I_B + |1\rangle\langle 1| \otimes W$ be two controlled unitaries from the A side, we have

$$U_3 = U_1 U U_2 = \begin{pmatrix} U_{00} & U_{01}W \\ VU_{10} & VU_{11}W \end{pmatrix}. \quad (2)$$

Since U is unitary, we have $U_{01}W = VU_{10}$. The matrix U_3 is a $2 \times d_B$ bipartite unitary of Schmidt rank at most 3, so it is a controlled unitary from the B side [15, 16]. We have proved that U is the product of three controlled unitaries U_1^\dagger, U_3 , and U_2^\dagger . There exist suitable local unitaries $S = I_A \otimes X_B$ and $T = I_A \otimes Y_B$, so that SU_3T is controlled in the computational basis of \mathcal{H}_B . Hence $U = (U_1^\dagger S^\dagger)(SU_3T)(T^\dagger U_2^\dagger)$ is a decomposition with each of the three parts controlled in the computational basis of \mathcal{H}_A or \mathcal{H}_B . Therefore $M = [(I_A \otimes E)(U_1^\dagger S^\dagger)](SU_3T)[(T^\dagger U_2^\dagger)(I_A \otimes F)]$ is exactly a 3-sandwich form. Hence the assertion holds.

(ii) From the proof of (i), we know that any $2 \times d_B$ unitary U has a 3-sandwich form. Let $U = V_1 V_2 V_3$ where V_1, V_3 are controlled unitaries controlled in the computational basis of \mathcal{H}_A , and V_2 is a controlled unitary controlled in the computational basis of \mathcal{H}_B . Since V_2 is controlled in the computational basis of \mathcal{H}_B , one can write $V_2 = \sum_{i,j=0}^1 |i\rangle\langle j| \otimes V_{ij}$ where all V_{ij} are diagonal matrices. By multiplying V_2 with two suitable diagonal controlled unitaries respectively from the left and right side, we can make all entries of V_{00}, V_{01} and V_{10} real and non-negative, and the entries of V_{11} real and non-positive. Since V_2 is unitary, we have $V_{00} = -V_{11}$ and $V_{01} = V_{10}$. So V_2 has Schmidt rank at most two. It is controlled from the A side [15]. The inverse of all diagonal unitary operators taken above are also diagonal, so they can be absorbed by V_1 and V_3 . The latter are still controlled unitaries from the A side in the computational basis. So $U = V_1 V_2 V_3$ is a 3-A form and the assertion holds.

(iii) The assertion follows from Lemma 1, which shows that the two-qubit SWAP gate is a product of three controlled unitaries, and no fewer. This completes the proof. \square

When $d_B = 2$ namely the unitary acts on two-qubit states, assertion (ii) has been proved as the statement that any two-qubit unitary has the so-called canonical form [19, 20]. It has been shown that any two-qubit unitary does not have Schmidt rank three [13]. For readers' reference, the Schmidt-rank-three multiqubit unitary has been investigated and constructed in [15, 17].

Now we are in a position to give an upper bound of $c(U)$ and the associated method of decomposing the bipartite unitary U .

Theorem 4 *Let U be a bipartite unitary on the $d_A \times d_B$ system. Then*

(i) *U has a $(2^{\lceil \log_2 d_A \rceil + 1} - 1)$ -sandwich form. Hence*

$$c(U) \leq 2^{\lceil \log_2 d_A \rceil + 1} - 1 \leq 4d_A - 5, \quad (3)$$

for any $d_A \geq 2$. In particular, $c(U) \leq 2d_A - 1$ when d_A is an integer power of 2.

(ii) *If all bipartite unitaries on the $d_A \times d_B$ system with odd $d_A \geq 3$ have $(2d_A - 1)$ -sandwich forms, then U has a $(2d_A - 1)$ -sandwich form for any even $d_A \geq 2$.*

Proof. (i) One can easily show that the second inequality in (3) holds. In particular its equality holds when $d_A = 2^n + 1$ with any nonnegative integer n . Since the first inequality in (3) and the last assertion of (i) both follow from the first assertion of (i), it is sufficient to prove the latter. The assertion is trivial if d_A or $d_B = 1$, so we assume $d_A, d_B \geq 2$. The proof is by induction over d_A . The assertion for $d_A = 2$ with any $d_B \geq 2$ is proven in Lemma 3. In the following we prove the assertion for a fixed $d_A \geq 3$, under the induction hypothesis that the $k \times d_B$ bipartite unitary with any $2 \leq k \leq d_A - 1$ and $d_B \geq 2$ has a $g(k)$ -sandwich form, where for any positive integer j we define

$$g(j) = 2^{\lceil \log_2 j \rceil + 1} - 1. \quad (4)$$

Let $\mathcal{H}_{A_1}, \mathcal{H}_{A_2} \subseteq \mathcal{H}_A$ be two subspaces spanned by the first y ($y \leq \lfloor d_A/2 \rfloor$) and $2y$ computational basis kets, respectively. Let $V = I_{A_1} \otimes I_B + V'$ be a BCU where V' is a bipartite unitary on the subspace $H = \mathcal{H}_{A_1}^\perp \otimes \mathcal{H}_B$. Let

$$W = W' + I_{A_2^\perp} \otimes I_B \quad (5)$$

be another BCU, where W' is a bipartite unitary on the subspace $\mathcal{H}_{A_2} \otimes \mathcal{H}_B$, and $I_{A_2^\perp}$ is the identity operator on the subspace $\mathcal{H}_{A_2}^\perp$. We can find a suitable V , such that in the top yd_B rows of the matrix product UV , the nonzero entries occur only in the first $2yd_B$ columns. Then we can find a suitable W such that the matrix product

$$X := UVW = I_{A_1} \otimes I_B + X', \quad (6)$$

where X' is a unitary acting on H . So X is a BCU controlled from the A side, and

$$U = XW^\dagger V^\dagger. \quad (7)$$

By regarding W' as a $2 \times yd_B$ bipartite unitary and using Lemma 3, we obtain that W' has a 3-sandwich form. Let

$$(W')^\dagger = CTD, \quad (8)$$

where C, D are both the direct sum of two unitaries each of order yd_B , and

$$T = \sum_{i=1}^{yd_B} W_i \otimes |i\rangle\langle i| \quad (9)$$

with some unitaries W_i of order two. So C, D and T can all be regarded as $2y \times d_B$ bipartite unitaries on the subspace $\mathcal{H}_{A_2} \otimes \mathcal{H}_B$. Using (5), (7), and (8), we have

$$U = X(CTD + I_{A_2^\perp} \otimes I_B)V^\dagger = (X\tilde{C})\tilde{T}(\tilde{D}V^\dagger), \quad (10)$$

where

$$\tilde{C} = C + I_{A_2^\perp} \otimes I_B, \quad (11)$$

$$\tilde{D} = D + I_{A_2^\perp} \otimes I_B, \quad (12)$$

$$\tilde{T} = T + I_{A_2^\perp} \otimes I_B. \quad (13)$$

It follows from (9) that T can be regarded as a controlled unitary on $\mathcal{H}_{A_2} \otimes \mathcal{H}_B$, controlled from the B side in the computational basis. This fact and (13) imply that \tilde{T} is a controlled unitary from the B side in the computational basis. Next, it follows from (6) and (11) that $X\tilde{C}$ is a BCU, i.e.,

$$X\tilde{C} = X_1 + X_2, \quad (14)$$

where the bipartite unitaries X_1 and X_2 act on the subspaces H^\perp and H , respectively. Since $\dim \mathcal{H}_{A_1} = y$ and $\dim \mathcal{H}_{A_1}^\perp = d_A - y$, they are both smaller than d_A for any $y = 1, 2, \dots, \lfloor d_A/2 \rfloor$. It follows from the induction hypothesis that X_1 and X_2 have $g(y)$ and $g(d_A - y)$ -sandwich forms, respectively. We have two decomposition

$$X_1 = \prod_{i=1}^{g(y)} X_{1,i}, \quad X_2 = \prod_{i=1}^{g(d_A-y)} X_{2,i}, \quad (15)$$

where for any odd and even i , the $X_{j,i}$ is a controlled unitary from the A and B side, respectively. Then so is $X_{1,i} + X_{2,i}$, because $X_{1,i}$ and $X_{2,i}$ act on the subspaces H^\perp and H , respectively. It follows from (4) and the condition $y \leq \lfloor d_A/2 \rfloor$ that $g(y) \leq g(d_A - y)$. This inequality, (14) and (15) imply $X\tilde{C} = \prod_{i=1}^{g(y)} (X_{1,i} + X_{2,i}) \cdot \prod_{j=g(y)+1}^{g(d_A-y)} (I_{A_1} \otimes I_B + X_{2,j})$. These facts imply that $X\tilde{C}$ has a $g(d_A - y)$ -sandwich form. Next using the same argument except that (11) is replaced by (12), one can show that $\tilde{D}V^\dagger$ also has a $g(d_A - y)$ -sandwich form. Third it

follows from (4) that $g(j)$ is odd for any positive integer j . Fourth in the paragraph below (13), we have shown that \tilde{T} is a controlled unitary from the B side in the computational basis. Applying these four facts to (10) implies that the unitary U has an x -sandwich form where

$$\begin{aligned} x &= \min_{1 \leq y \leq \lfloor d_A/2 \rfloor} (2g(d_A - y) + 1) \\ &= 2g(\lceil d_A/2 \rceil) + 1 = g(d_A). \end{aligned} \quad (16)$$

The last two equalities in (16) follow from (4), and the fact that $\lceil \log_2 d_A \rceil = \lceil \log_2(d_A + 1) \rceil$ for odd $d_A \geq 3$. So (16) is exactly the first assertion of (i).

(ii) The proof is by induction over even $d_A \geq 2$. The assertion for $d_A = 2$ with any $d_B \geq 2$ is proven in Lemma 3. In the following we prove the assertion for a fixed even $d_A \geq 4$, under the induction hypothesis that the $k \times d_B$ bipartite unitary with any even $k \in [2, d_A - 1]$ and $d_B \geq 2$ has a $(2k - 1)$ -sandwich form. One can verify that the argument from the paragraph below (4) to the second sentence below (14) still applies here. We choose $y = d_A/2$ in the argument. If y is odd (respectively, even), then the condition in (ii) (respectively, the induction hypothesis) implies that X_1 and X_2 in (14) both have $(d_A - 1)$ -sandwich forms, respectively. Hence (15) and the subsequent paragraph hold, except that $g(j)$ is replaced by $2j - 1$ for any positive integer j . Since $d_A \geq 2$ is even, applying these facts to (10) implies that the unitary U has an x -sandwich form where

$$x = 2(d_A - 1) + 1 = 2d_A - 1. \quad (17)$$

This completes the proof of assertion (ii). \square

We do not know whether the condition in Theorem 4 (ii) can be satisfied, and we leave it as an open problem. As a byproduct of the theorem, it follows from (7) that

Corollary 5 *Any bipartite unitary is the product of three BCUs controlled from the A , B and A sides, respectively.*

It is known that any two-qubit BCU is a controlled unitary. Hence Lemma 3 (iii) implies that the two-qubit CNOT gate cannot be the product of only two BCUs. In other word, the upper bound three in Corollary 5 is tight.

The upper bound obtained in Theorem 4 is $4d_A - 5$ and it is polynomially smaller than $4d_A^4$ obtained in [2]. Compared to the latter, the implementation of a bipartite unitary by arbitrary controlled unitaries can indeed save quantum resources. Since the systems A and B are symmetric in the problem, $4d_B - 5$ is also an upper bound for the number of controlled gates. We consider the optimality of the bound $4d_A - 5$ under the assumptions that $d_A \leq d_B$ and that the number of controlled gates is a function of d_A only. By parameter counting, the $4d_A - 5$ is already optimal up to a constant factor, because the entire unitary has $d_A^2 d_B^2$ free real parameters in it, and each controlled unitary from the A side and controlled in the computational basis of \mathcal{H}_A has $d_A d_B^2$ free real parameters in it, while each controlled gate from the B side has

$d_B d_A^2$ free real parameters in it, less than what is in a controlled gate from the A side (so that a larger number of these would be used if they are used instead of controlled gates from the A side). Note that for two adjacent controlled gates, we have overestimated the number of free parameters, since when they are both controlled from the A side, the change of controlling basis on \mathcal{H}_A could be viewed as a change in either of the controlled gates, and generally, a bipartite diagonal gate between two adjacent controlled gates can be absorbed into any of the two adjacent controlled gates. But such issues only affect the count above by a lower order factor.

We comment on the connection with the results in the literature. Our Lemma 3(i) in the special case that d_B is an integer power of 2 is the same as Theorem 10 of Shende *et al.* [8] (see also [21]). Our Theorem 4(i) in the case that d_A is an integer power of 2 can also be derived by recursively applying Theorem 10 of [8] (the first step of recursion is illustrated in Theorem 11 of [8], and note that a gate controlled by multiple qubits belonging to the same party is a controlled gate in our language). We abbreviate the details here. Therefore our result can be viewed as a generalization of the results in [8] to the general dimensions. Based on our result, it may be possible to decompose any qudit circuit (with dimensions of qudits not required to be all equal) using controlled two-qudit unitaries. The following Sec. IV can be viewed as a step in this direction, but we do not decompose the gates fully there, allowing some gate controlled by multiple qudits. There may be some extensions of the techniques in [8] to the case of higher dimensional qudits that can help decompose such multiply-controlled gate. There are some papers on decomposition of qudit circuits, such as [3, 6, 7]. It is possible that the methods in those papers may be combined with the results in this paper to give a better upper bound of the number of two-qudit (controlled) gates needed. Apart from the application to circuit decomposition, the other potential application is to help study the nonlocal resource usage in implementing nonlocal unitaries. Here the usage of nonlocal resources is to be optimized, and the local resources such as local unitaries are deemed as cheap. Section V is a step in this direction, but it only discusses the cost in terms of a particular type of nonlocal gate (whose implementation cost is upper bounded by a constant), and not in terms of the more conventional resources such as entanglement.

A. Decomposition of complex permutation matrices

The upper bound in Theorem 4 works for arbitrary bipartite unitaries, and it increases linearly with the dimension. One may expect to have a constant upper bound for some special bipartite unitaries. In this subsection we give such a bound for any *complex permutation matrix* in Theorem 7. It is a unitary matrix with one and only one nonzero element on each row and column. When the

nonzero elements have no phases and are equal to one, it becomes the standard permutation matrix. The complex permutation matrix is mathematically known as a special monomial matrix, and has been used to characterize the mutually unbiased bases [22, 23]. The complex permutation gate is of interest to the study of quantum computation, as it is a somewhat classical part of a quantum circuit; see its use in the definition of the Fourier hierarchy in [9]. The diagonal unitary, which is a special complex permutation matrix, can be efficiently simulated in terms of the Clifford+T basis by the algorithm in [24]. We define the *controlled-permutation matrices* to be bipartite controlled unitaries controlled in the computational basis of one system and with the terms on the controlled side being permutation matrices. The *controlled-complex-permutation matrices* are defined similarly.

To study the decomposition of complex permutation matrices, we present a preliminary lemma, which is actually a form of the Hall's marriage theorem [25]. Suppose $V = \sum_{j,k=1}^{d_A} |j\rangle\langle k| \otimes V_{j,k}$ is a bipartite operator on the space $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$. We say that V is *absolutely singular* if there are integers j_1, \dots, j_s and k_1, \dots, k_t with $s+t > d_A$, such that $V_{j_a, k_b} = 0$. The absolute singularity of V is unchanged up to any product permutation operators on the left- and right-hand sides of V (a product permutation operator is of the form $P_A \otimes Q_B$, where P_A and Q_B are local permutation operators; in what follows we only need Q_B to be an identity matrix). Hence an absolute singular V is locally equivalent to another bipartite operator whose left-upper $s d_B \times t d_B$ submatrix is zero. Evidently an absolutely singular operator is singular, but the converse is not true. We characterize the absolute singularity as follows.

Lemma 6 $V = \sum_{j,k=1}^{d_A} |j\rangle\langle k| \otimes V_{j,k}$ is not absolutely singular if and only if there are d_A distinct integers k_1, \dots, k_{d_A} , such that the blocks $V_{1,k_1}, \dots, V_{d_A, k_{d_A}}$ are all nonzero.

Proof. We first present a matrix-based proof, and then provide a proof of the equivalence of the lemma to Hall's marriage theorem, which is known to have several different proofs.

Matrix-based proof. The “if” part follows from the definition of absolute singularity. Let us prove the assertion in the “only if” part. Assume V is not absolutely singular. This assumption and the assertion are both unchanged up to any product permutation operators on the left- and right-hand sides of V . We will refer to the $d_B \times d_B$ blocks in V still as $V_{j,k}$ since there is no confusion. The assertion is trivial for $d_A = 1, 2$. Next we shall use induction over d_A . The induction hypothesis is that the assertion holds when d_A is replaced by $2, \dots, d_A - 1$, and we will prove the assertion for d_A . Since V is not absolutely singular, we may assume that V_{11} is nonzero up to a suitable product permutation operator on the right-hand side of V . If the submatrix $X = \sum_{j,k=2}^{d_A} |j\rangle\langle k| \otimes V_{j,k}$ is not

absolutely singular, then the assertion follows from the induction hypothesis on X . Suppose X is absolutely singular. By performing two suitable product permutation operators, respectively, from the left- and right-hand side of V , we may assume that $V_{j,k} = 0$ where $j = 2, \dots, s$, $k = t+1, \dots, d_A$, and $d_A \geq s > t \geq 1$. Since V is not absolutely singular, we have $s = t+1$. Using a suitable product permutation operator on the left-hand side of V , we may assume that $V = \begin{pmatrix} V_1 & 0 \\ V_2 & V_3 \end{pmatrix}$, where V_1 and V_3 are, respectively, $(s-1)d_B \times (s-1)d_B$ and $(d_A - s + 1)d_B \times (d_A - s + 1)d_B$ submatrices. Since V is not absolutely singular, neither are V_1 and V_3 . The hypothesis induction implies that the assertion holds for both V_1 and V_3 . Hence the assertion holds for V . This completes the proof.

Equivalence of the lemma to Hall's marriage theorem. We use the combinatorial formulation of Hall's marriage theorem in [25]. It involves some given elements, each of which may be in one or more of some given sets. There is a *marriage condition* that says the number of distinct elements contained in k sets is at least k , for any integer $k \geq 0$. A *system of distinct representatives* is a set of distinct elements, each of which is in a different set. Hall's marriage theorem says that a system of distinct representatives exists if and only if the marriage condition is satisfied. Let us now describe the equivalence of the current lemma to the above theorem. Take the sets to be the big rows of V labeled by j , and the elements to be the big columns labeled by k , and let an element k be in a set j if and only if the $V_{j,k}$ is nonzero. Then the marriage condition corresponds to the definition of absolute singularity, and a system of distinct representatives corresponds to a sequence of d_A distinct big column labels k_i ($i = 1, \dots, d_A$) such that V_{i, k_i} is nonzero. This establishes the equivalence. \square

Theorem 7 Any bipartite complex permutation unitary has a 3-sandwich form, composed of controlled-complex-permutation matrices. In particular, if the unitary is a permutation matrix, the 3-sandwich form is composed of controlled-permutation matrices.

Proof. The second claim implies the first claim, since any complex permutation unitary is the product of a permutation matrix and a diagonal unitary, the latter can be absorbed into one of the controlled-permutation matrices in the decomposition of the complex permutation unitary. Therefore it suffices to prove the second claim. Suppose U is a bipartite permutation unitary on the $d_A \times d_B$ system.

Let $U = \sum_{j,k=1}^{d_A} |j\rangle\langle k| \otimes U_{j,k}$. Since it is not absolutely singular, it follows from Lemma 6 that there are d_A distinct integers k_1, \dots, k_{d_A} , such that the blocks $U_{1, k_1}, \dots, U_{d_A, k_{d_A}}$ are all nonzero. There are two controlled-permutation matrices $V = \sum_{j=1}^{d_A} |j\rangle\langle j| \otimes V_j$ and $W = \sum_{j=1}^{d_A} |j\rangle\langle j| \otimes W_j$ from the A side, such that the first entry of any one of the blocks

$V_1 U_{1,k_1} W_{k_1}, \dots, V_{d_A} U_{d_A, k_{d_A}} W_{k_{d_A}}$ of VUW is one. If $d_B = 2$ then VUW is a controlled-permutation unitary from the B side. So the assertion holds. We use the induction over $d_B \geq 2$. We have $VUW = X \otimes |1\rangle\langle 1| + Y$, where X is a permutation matrix on \mathcal{H}_A , and Y is a permutation matrix on $\mathcal{H}_A \otimes |1\rangle^\perp$. The induction hypothesis on Y implies that $Y = Y_1 Y_2 Y_3$, where Y_1, Y_2 and Y_3 are controlled-permutation matrices from A, B and A side, respectively. Hence

$$U = V^\dagger (I_A \otimes |1\rangle\langle 1| + Y_1) (X \otimes |1\rangle\langle 1| + Y_2) \cdot (I_A \otimes |1\rangle\langle 1| + Y_3) W^\dagger, \quad (18)$$

which is a 3-sandwich form of U composed of controlled-permutation matrices. So we have proved the second claim. This completes the proof. \square

It is known that the SWAP_d gate defined in Lemma 1 has a decomposition using three bipartite controlled gates [18]. In agreement with the construction in [18], Theorem 7 shows that the three gates can be chosen as controlled-permutation gates in a 3-sandwich form.

The theorem also has implications for classical circuits. Define a *classical reversible circuit (classical permutation gate)* to be a classical circuit that is a permutation on the allowed set of input data. In the bipartite case, suppose d_A and d_B are the number of possible states on the systems A and B , respectively, then we say the circuit acts on a $d_A \times d_B$ system. For example, when n_A and n_B are the number of bits on the two systems, we have $d_A = 2^{n_A}$ and $d_B = 2^{n_B}$. From Theorem 7, and noting that in the proof of Theorem 7 there is no requirement of coherence in both the target unitary and the controlled unitaries in the decomposition, we have

Corollary 8 *Suppose T is a classical reversible circuit on a $d_A \times d_B$ bipartite system, then T can be implemented using the product of 3 bipartite classical controlled-permutation gates.*

Note the classical controlled-permutation gates are controlled in the computational basis, as one would expect. Corollary 8 is also stated in Sec. 3.15 and Appendix E of [26], where the proof approach is by considering the permutation accomplished by the circuit and directly using the Birkhoff-von Neumann theorem (explained below), which has an integer-arithmetic version that says the following: Any matrix of size $n \times n$ with non-negative integer entries and with row and column sums equal to q can be decomposed as the sum of q permutation matrices of size $n \times n$. Such a statement appears in [27], and a simple proof is by repeated use of Hall's marriage theorem, each time finding a permutation matrix, which is to be subtracted from the original matrix, and this process terminates when the resulting matrix becomes the zero matrix. The construction of the 3 classical permutation gates is as follows: arrange the $d_A \times d_B$ computational-basis states in a rectangular table with d_A rows and d_B columns, and define a matrix M to contain integer elements M_{ij} that indicate how many elements in row i of

the table are to be transferred to row j of the table after the permutation gate T . The first, second, and third controlled-permutation gates permutes among elements in the same row, column, and row, respectively. Each column of the rectangular table after the first gate contains elements that are to be permuted under the second gate. Each permutation in a column corresponds to one permutation matrix in the decomposition of M as the sum of permutation matrices. The argument above roughly describes the proof in [26] for Corollary 8. In comparison, our matrix-based approach for obtaining the circuit decomposition hints at some connections to the sandwich form of general unitaries of the sort in Lemma 3 and Theorem 4.

IV. DECOMPOSITION OF MULTIPARTITE UNITARY OPERATORS

In this section, we study the decomposition of n -partite unitary operators U on the space $\otimes_{j=1}^n \mathcal{H}_j$ with $\text{Dim } \mathcal{H}_i = d_i$. We define a *generalized m -sandwich form* of U to be a decomposition of the form $U = U_1 U_2 \cdots U_m$, where any U_i is a controlled unitary controlled in the computational basis from $n - 1$ fixed systems. For example, U_1 may be controlled from the systems of $\mathcal{H}_1, \dots, \mathcal{H}_{n-1}$, U_2 may be controlled from the systems of $\mathcal{H}_1, \dots, \mathcal{H}_{n-2}, \mathcal{H}_n$, etc. The computational basis in $\otimes_{j=1}^n \mathcal{H}_j$ consists of the product states $|j_1, \dots, j_n\rangle$ where $j_i = 1, \dots, d_i$ for each i . The word “fixed” means the choices of controlling parties are fixed for each gate U_i . Such choices are a function of the generalized m -sandwich form that we choose. In the results in this section, we always fix such choices. We have

Proposition 9 *Any n -partite unitary has a generalized $[2 \prod_{j=1}^{n-1} (2d_j - 2) - 1]$ -sandwich form.*

Proof. Let $f(n) = 2 \prod_{j=1}^{n-1} (2d_j - 2) - 1$. The assertion is trivial for $n = 1$, and follows from Theorem 4 for $n = 2$. We use the induction on n . Assume that any $(n - 1)$ -partite unitary has a generalized $f(n - 1)$ -sandwich form. Let U be an n -partite unitary. By regarding $\mathcal{H}_A = \mathcal{H}_1$ and $\mathcal{H}_B = \otimes_{j=2}^n \mathcal{H}_j$ in Theorem 4, we obtain the $(4d_1 - 5)$ -sandwich form

$$U = \prod_{j=1}^{4d_1-5} U_j, \quad (19)$$

where U_j is controlled in the computational basis of \mathcal{H}_A for odd j , and of \mathcal{H}_B for even j , respectively. In particular, the computational basis in the latter is realized by performing suitable unitaries on \mathcal{H}_B that can be absorbed by the U_j with odd j . Then

$$U_j = \bigoplus_{k=1}^{d_1} |k\rangle\langle k| \otimes U_{jk}, \quad \forall \text{ odd } j, \quad (20)$$

where each U_{jk} is a unitary on \mathcal{H}_B . From the induction assumption, U_{jk} has a generalized $[2 \prod_{j=2}^{n-1} (2d_j - 2) - 1]$ -sandwich form. Then (20) implies that U_j with any odd j has a generalized $[2 \prod_{j=2}^{n-1} (2d_j - 2) - 1]$ -sandwich form. Since U_j with any even j is a controlled unitary controlled in the computational basis of \mathcal{H}_B , (19) implies that U has a generalized m -sandwich form where

$$\begin{aligned} m &= (2d_1 - 2)[2 \prod_{j=2}^{n-1} (2d_j - 2) - 1] + 2d_1 - 3 \\ &= f(n). \end{aligned} \quad (21)$$

This completes the proof. \square

The proof above first divides the systems into two groups of one party and $(n - 1)$ parties each. When $n \geq 4$, there are also other ways of dividing the systems at the first step that may give rise to fewer gates in the generalized sandwich form. The following result is for the case of $n = 4$.

Proposition 10 *Any unitary on four parties A, B, C, D has a generalized $[4(d_A d_B - 1)(2d_A + 2d_C - 5) - 4d_A + 5]$ -sandwich form.*

Proof. Let U be a unitary on these four parties. By regarding \mathcal{H}_A and \mathcal{H}_B in Theorem 4 as \mathcal{H}_{AB} and \mathcal{H}_{CD} respectively, we obtain the following sandwich form

$$U = \prod_{j=1}^{4d_A d_B - 5} U_j, \quad (22)$$

where U_j is controlled in the computational basis of \mathcal{H}_{AB} for odd j , and in the computational basis of \mathcal{H}_{CD} for even j , respectively. Then

$$U_j = \bigoplus_{k=1}^{d_A d_B} |k\rangle\langle k|_{AB} \otimes U_{jk}, \quad \forall \text{ odd } j, \quad (23)$$

where $|k\rangle\langle k|_{AB}$ are projectors onto the computational basis of \mathcal{H}_{AB} , and each U_{jk} is a unitary on \mathcal{H}_{CD} . From Theorem 4, U_{jk} has a generalized $(4d_C - 5)$ -sandwich form. Then (23) implies that U_j with any odd j has a generalized $(4d_C - 5)$ -sandwich form. Similarly, U_j with any even j has a generalized $(4d_A - 5)$ -sandwich form. Therefore U is the product of

$$\begin{aligned} &(2d_A d_B - 2)(4d_C - 5) + (2d_A d_B - 3)(4d_A - 5) \\ &= 4(d_A d_B - 1)(2d_A + 2d_C - 5) - 4d_A + 5 \end{aligned} \quad (24)$$

unitaries that are controlled in the computational basis of 3 parties. This completes the proof. \square

To compare the two Propositions above, assume $n = 4$ in Proposition 9 with the subscripts 1, 2, 3, 4 replaced by A, B, C, D , respectively, and that $d_A \leq d_B \leq d_C \leq d_D$. Then Proposition 9 gives that U is the product of $16(d_A - 1)(d_B - 1)(d_C - 1) - 1$ unitaries that are controlled from

3 parties. Therefore, at least when $d_B \ll d_C$ and d_A is a large constant (say $d_A \geq 20$), Proposition 10 gives a smaller number than Proposition 9.

The proofs of the results above imply that, if we could reduce the number of bipartite controlled unitaries in the sandwich form in Theorem 4, then the number of multipartite controlled unitaries in the generalized sandwich form could also be reduced. In particular, from Theorem 7, we have

Proposition 11 *Any n -partite complex permutation unitary has a generalized $(2n - 1)$ -sandwich form composed of controlled-complex-permutation unitaries controlled by $n - 1$ parties.*

Proof. It suffices to consider permutation unitaries, for the same reason as stated in the proof of Theorem 7. From Theorem 7, the claim holds for $n = 2$. The proof is by induction over n . The induction hypothesis is that the claim holds when n is replaced by any positive integer less than n . Now consider $n \geq 3$, and take a bipartite cut of the first $n - 1$ parties versus the last party. From Theorem 7, the permutation unitary has a 3-sandwich form, and the first and the last gates in the 3-sandwich form are a controlled permutation controlled from the first $n - 1$ parties. The middle gate in the 3-sandwich form is a controlled permutation controlled from the last party, so it is of the form $U_1 \otimes |1\rangle\langle 1| + U_2 |2\rangle\langle 2|$, where the permutation operators U_1 and U_2 on the first $n - 1$ parties can each be decomposed into $2(n - 1) - 1$ controlled-permutation gates controlled by $n - 2$ parties, and the choices of those controlling $n - 2$ parties are always the same for the decompositions of U_1 and U_2 , according to the induction hypothesis. Therefore the permutation unitary on n parties has a generalized $(2n - 1)$ -sandwich form composed of controlled-permutation gates controlled by $n - 1$ parties. The case with phases is similar, just adding the word “complex”. This completes the proof. \square

The result above has a corresponding statement for classical reversible circuits. In the special case that each party is one bit, it is illustrated by a sample circuit in Fig. 2 of [28] (note the sequence of lines is opposite from that in the proof above).

As mentioned in Sec. III, it is possible that the literature results on the decomposition of qudit circuits [3, 6, 7] could be combined with the results in this section to give better upper bounds of the number of two-qudit gates.

V. DECOMPOSITION USING A SIMPLE TYPE OF GATES

In this section, we apply our result on decomposition using controlled unitaries to the decomposition using more basic type of gates defined below. One of our motivations is to characterize the nonlocal part of the cost for implementing bipartite unitaries using some measure

with a fixed unit, rather than using the number of controlled unitaries which is a measure with its unit dependent on the dimensions. The cost measure that we use is the number of standard gates defined below, and we do not allow any ancillary systems in the circuit. The case with ancillas will be discussed in Sec. VII. In the following definitions, I_X stands for the identity operator on system X .

Definition 12 A standard gate is a unitary acting on the Hilbert space $\mathcal{H}_{AB} = \mathcal{H}_A \otimes \mathcal{H}_B$ of the form $U = U_{AB} = (V_{ab} \oplus I_{AB \setminus ab})$, where $\mathcal{H}_{AB} = \mathcal{H}_{ab} \oplus \mathcal{H}_{AB \setminus ab}$, and $\mathcal{H}_a \subseteq \mathcal{H}_A$ and $\mathcal{H}_b \subseteq \mathcal{H}_B$ are two-dimensional each, and V_{ab} is a Schmidt-rank-2 unitary on the 2×2 space $\mathcal{H}_{ab} = \mathcal{H}_a \otimes \mathcal{H}_b$. The V_{ab} is called the nontrivial part of U .

Note the word ‘‘Schmidt-rank-2’’ above can be replaced by ‘‘controlled’’, as Schmidt-rank-2 unitaries are controlled unitaries ([15]; also see an alternative proof in [17]), and two-qubit unitaries of Schmidt rank greater than 2 must have Schmidt rank 4 [13] and thus cannot be controlled unitaries. The case of \mathcal{H}_{AB} being strictly larger than \mathcal{H}_{ab} is useful, for example, in the decomposition of the Toffoli gate [29], and has been experimentally realized [30]. The definition above can be extended to a more general definition below:

Definition 13 A bipartite elementary gate is a unitary acting on the Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B = (\mathcal{H}_a \otimes \mathcal{H}_C \oplus \mathcal{H}_D) \otimes (\mathcal{H}_b \otimes \mathcal{H}_E \oplus \mathcal{H}_F)$ of the form $U = (V_{ab} \otimes I_{CE}) \oplus I_{AB \setminus abCE}$, where \mathcal{H}_a and \mathcal{H}_b are two-dimensional each, and V_{ab} is a Schmidt-rank-2 unitary, and $\mathcal{H}_{AB} = \mathcal{H}_{AB \setminus abCE} \oplus \mathcal{H}_{abCE}$.

In the following we consider the decomposition of bipartite unitary operators into the product of bipartite standard gates defined in Definition 12 and arbitrary local gates, with the goal of minimizing the number of non-local standard gates. The more general Definition 13 will not be studied in this paper except that we define some gate cost using it in Definition 14 and raise some open questions.

We define the following gate-cost measures for a bipartite unitary.

Definition 14 Let $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$ be the complex Hilbert space of a finite-dimensional bipartite quantum system, with $\text{Dim } \mathcal{H}_A = d_A$ and $\text{Dim } \mathcal{H}_B = d_B$. For any given bipartite unitary $U : \mathcal{H} \rightarrow \mathcal{H}$,

$$\begin{aligned} c_s(U) &:= \min\{k | U = U_1 U_2 \cdots U_k, \quad U_i \in \mathcal{S}_s\}, \\ c_e(U) &:= \min\{k | U = U_1 U_2 \cdots U_k, \quad U_i \in \mathcal{S}_e\}, \end{aligned} \quad (25)$$

where \mathcal{S}_s (respectively, \mathcal{S}_e) is the set of bipartite unitaries on the same space that are equivalent to the standard (respectively, bipartite elementary) gates under local unitaries.

In the case $d_A = d_B = 2$, it is well known that three Schmidt-rank-2 gates are sufficient and necessary for a general two-qubit unitary [12], as mentioned in Sec. III. An example that needs three Schmidt-rank-2 gates is the two-qubit SWAP gate ([12], also see Lemma 1). Our main result for general $d_A \times d_B$ system is as follows.

Proposition 15 (i) Any bipartite unitary on $d_A \times d_B$ system is the product of $f(d_A, d_B)$ standard gates interspersed with local unitaries on \mathcal{H}_A or \mathcal{H}_B , where

$$\begin{aligned} f(d_A, d_B) &= 2(d_A - 1)^2 \lfloor \frac{d_B}{2} \rfloor \\ &+ (2d_A - 3)(d_B - 1) \lfloor \frac{d_A}{2} \rfloor. \end{aligned} \quad (26)$$

(ii) If the unitary is a controlled unitary controlled from the A side, then

$$f(d_A, d_B) = (d_A - 1) \lfloor \frac{d_B}{2} \rfloor. \quad (27)$$

(iii) If the unitary is a complex permutation unitary, then

$$f(d_A, d_B) = 2(d_A - 1) \lfloor \frac{d_B}{2} \rfloor + (d_B - 1) \lfloor \frac{d_A}{2} \rfloor. \quad (28)$$

(iv) If the nontrivial part of the standard gates is required to be CNOT, then at most $3(d_A - 1)(d_B - 1)$ such standard gates together with local permutation gates can implement any bipartite permutation unitary on $d_A \times d_B$ space.

Proof. (i). Let U be the bipartite unitary. Theorem 4 implies that U has the following sandwich form

$$U = \prod_{j=1}^{4d_A-5} U_j, \quad (29)$$

where U_j is controlled in the computational basis of \mathcal{H}_A for odd j , and in the computational basis of \mathcal{H}_B for even j , respectively. For all odd j , we have

$$\begin{aligned} U_j &= \bigoplus_{k=1}^{d_A} |k\rangle\langle k|_A \otimes U_{jk}, \\ &= \prod_{k=1}^{d_A} [|k\rangle\langle k|_A \otimes U_{jk} \oplus (I_A - |k\rangle\langle k|_A) \otimes I_B], \end{aligned} \quad (30)$$

where $|k\rangle\langle k|_A$ are projectors onto the computational basis of \mathcal{H}_A , and each U_{jk} is a unitary on \mathcal{H}_B . We can apply a local unitary U_{jd_B} on \mathcal{H}_B before performing other steps below. In order to implement U_j , the operator that remains to be implemented is still given by (30) but with U_{jd_B} becoming the identity matrix, and the other operators U_{jk} also changed but we still denote the changed matrices as U_{jk} , with $1 \leq k \leq d_A - 1$. The U_j is to be implemented using the product of $d_A - 1$ operators, as shown in the second line of (30). Then each of the U_{jk} with $1 \leq k \leq d_A - 1$ can be assumed to be a diagonal

unitary, because we can apply a suitable local unitary similarity transform on \mathcal{H}_B so that U_{jk} is diagonal and I_B is unchanged. By a local diagonal unitary gate on \mathcal{H}_A which only applies a phase on $|k\rangle_A$, we can set the last diagonal element of U_{jk} to be 1, while the I_B corresponding to the basis kets in \mathcal{H}_A other than $|k\rangle_A$ are unchanged. Therefore we have

$$U_{jk} = \text{diag}(x_1^{(jk)}, x_2^{(jk)}, \dots, x_{d_B-1}^{(jk)}, 1), \quad (31)$$

where $x_i^{(jk)}$ are complex phases, $i = 1, 2, \dots, d_B - 1$. Then we choose $\lfloor \frac{d_B}{2} \rfloor$ standard gates as follows:

$$V_r^{(jk)} = I_A \otimes I_B + (x_{2r-1}^{(jk)} - 1)|k\rangle\langle k|_A \otimes |2r-1\rangle\langle 2r-1|_B \\ + (x_{2r}^{(jk)} - 1)|k\rangle\langle k|_A \otimes |2r\rangle\langle 2r|_B, \text{ for } 1 \leq r \leq \lfloor \frac{d_B}{2} \rfloor, \quad (32)$$

Each gate $V_r^{(jk)}$ applies phases on the two states $|k\rangle_A \otimes |2r-1\rangle_B$ and $|k\rangle_A \otimes |2r\rangle_B$, but keeps other computational basis states of \mathcal{H}_{AB} unchanged. It is easy to verify that such a gate has Schmidt rank at most 2 when viewed as a unitary acting on the 2×2 system with basis $\{|k'\rangle_A, |k\rangle_A\} \times \{|2r-1\rangle_B, |2r\rangle_B\}$, where $k' \neq k$. Hence for each (j, k) pair with odd j and $1 \leq k \leq d_A - 1$, we need $\lfloor \frac{d_B}{2} \rfloor$ standard gates to implement the operator $|k\rangle\langle k|_A \otimes U_{jk} \oplus (I_A - |k\rangle\langle k|_A) \otimes I_B$ in the last line of (30). Therefore, for each odd j , U_j needs $(d_A - 1)\lfloor \frac{d_B}{2} \rfloor$ standard gates to implement, assisted by local unitaries. Similarly, for each even j , U_j needs $(d_B - 1)\lfloor \frac{d_A}{2} \rfloor$ standard gates to implement, assisted by local unitaries. The assertion then follows by counting the numbers of U_j in (29) in terms of odd and even j . This completes the proof of (i).

(ii). The claim follows from the proof of (i) by setting the upper bound for j in (29) to 1.

(iii). The claim follows from Theorem 7 and the result of (ii) applied to the A and B sides.

(iv). From Theorem 7, every bipartite permutation unitary is the product of 3 controlled-permutation unitaries, controlled from the A , B and A side, respectively. Every permutation on n elements is the product of at most $n - 1$ transpositions (swap of two elements). Define a controlled-transposition gate to be a bipartite unitary of the form $|1\rangle\langle 1|_A \otimes I_B + |2\rangle\langle 2|_A \otimes V_B$, where $V_B = |j\rangle\langle k| + |k\rangle\langle j| + \sum_{i \neq j, k} |i\rangle\langle i|$, for some $j \neq k$ ($\{|i\rangle\}$ is the computational basis of \mathcal{H}_B). For the special case $d_A = 2$, up to a local permutation on \mathcal{H}_B we can write a controlled-permutation gate from the A side as $|1\rangle\langle 1| \otimes I_B + |2\rangle\langle 2| \otimes P_2$, where P_2 is a permutation unitary on \mathcal{H}_B . This controlled-permutation gate can be written as the product of at most $d_B - 1$ controlled-transposition gates, which are standard gates with their nontrivial part being the CNOT. For larger d_A , up to a local permutation on \mathcal{H}_B we can write the controlled-permutation gate as $|1\rangle\langle 1| \otimes I_B + \sum_j |j\rangle\langle j| \otimes P_j$, where P_j are permutation unitaries on \mathcal{H}_B . Take the subspace $\text{span}\{|1\rangle_A, |j\rangle_A\}$ ($2 \leq j \leq d_A$) as the A side space in the $d_A = 2$ result above; we have that at most $d_B - 1$

standard gates with the nontrivial part being CNOT can implement $(I_A - |j\rangle\langle j|) \otimes I_B + |j\rangle\langle j| \otimes P_j$. Repeat this $d_A - 1$ times for $j = 2, \dots, d_A$, a controlled-permutation gate from the A side can be implemented using at most $(d_A - 1)(d_B - 1)$ such standard gates. The last result is the same for the B side. Hence the claim follows. \square

It can be verified that in Proposition 15(iv), the phrase “together with local permutation gates” can be dropped by allowing the nonlocal unitary to be implemented up to local permutations before and after it. Since a permutation gate on d -dimensional space requires at most $d - 1$ transpositions of the type $|j\rangle\langle k| + |k\rangle\langle j| + \sum_{i \neq j, k} |i\rangle\langle i|$, the four local permutations on \mathcal{H}_A or \mathcal{H}_B require at most $2d_A + 2d_B - 4$ local transpositions in total. Therefore the total number of standard gates of the CNOT type and the local transpositions is at most $3(d_A - 1)(d_B - 1) + 2d_A + 2d_B - 4 = 3d_A d_B - d_A - d_B - 1$. It could potentially be further reduced by a constant factor, and this is listed as an open problem in the Conclusions.

From [18] and Proposition 15 (ii), the SWAP_d gate has a decomposition using $3(d - 1)\lfloor \frac{d}{2} \rfloor$ standard gates across the two systems, together with some local unitaries. On the other hand, if we are not restricted to writing the SWAP_d gate as a product of some gates, but consider the actual cost of implementation, we could also make use of tensor products. Suppose $d = \prod_{j=1}^m p_j$, where $m \geq 1$ is an integer and p_j are primes. Then the SWAP_d gate is the tensor product of the SWAP gates on $p_j \times p_j$ systems. The SWAP gate on $p_j \times p_j$ system has a decomposition using $3(p_j - 1)\lfloor \frac{p_j}{2} \rfloor$ bipartite standard gates, together with some local unitaries. Hence the total implementation cost is $\sum_{j=1}^m 3(p_j - 1)\lfloor \frac{p_j}{2} \rfloor$ bipartite standard gates, together with some local unitaries.

VI. THE ROLE OF SCHMIDT RANK IN DECOMPOSITION OF BIPARTITE UNITARIES

The Schmidt rank of a bipartite unitary U sometimes determines the number of bipartite controlled unitaries needed to decompose U , as it is proved in [15, 17] that $c(U) = 1$ when $\text{Sch}(U) = 2$ or 3. To investigate the relation between $c(U)$ and $\text{Sch}(U)$ for general bipartite unitary U , we discuss the different cases characterized by how large $r := \text{Sch}(U)$ is compared to the dimensions d_A and d_B . If $r \geq \min\{d_A, d_B\}$, then it follows from Theorem 4 (applied to the A or B side) that $c(U) \leq 4r - 5$. On the other hand if $r < \min\{d_A, d_B\}$, then we need to count the number of parameters in U . It is equal to $(d_A^2 - r + d_B^2)r$, which is smaller than $2d_A d_B^2$ when $d_A \leq d_B$. A controlled unitary from the A side contains $d_A d_B^2$ parameters, and noting that there are some redundant parameters when counting consecutive controlled unitaries in a product, theoretically U could be the product of only three controlled unitaries (or even two when r is further restricted to smaller values). But the actual number may be higher. A possible class of candidate examples that *may* need more than three con-

trolled unitaries is the U' in Example 16 below.

We now show a class of examples where $c(U)$ is much smaller than $\text{Sch}(U)$ (note that a generic permutation matrix already has this property, according to Theorem 7, but our interest here is to show the derived class of examples U' that fit into the requirement $r < \min\{d_A, d_B\}$ in the previous paragraph).

Example 16 Let V_{CB} be a generic unitary on $d \times d$ system of Schmidt rank d^2 with $d > 2$, and let $U = V_{CB} \otimes I_D$, where D is of the same size as B and C (d dimensions). Then U is of Schmidt rank d^2 across the bipartite cut CD - B . But there is a decomposition using only 6 controlled unitaries: first, swap the states of the systems D and B , using 3 controlled gates [18], then do the V on CD , and finally swap the D and B again, using another 3 controlled gates. The local unitary on CD in the second step could be absorbed into the two controlled unitaries before and after it, thus only 6 controlled unitaries are needed in total without extra local unitaries. Now consider the unitary $U' = U \otimes I_E \otimes I_F$, where E is one qubit and F is of dimension $2d$. Then U' of Schmidt rank d^2 across the CDE - BF cut, and $c(U') \leq 6$, and the Schmidt rank $r = d^2$ satisfies $r < \min\{d_{CDE}, d_{BF}\}$, fitting into the requirement in the first paragraph of this section.

Speaking about the general dependence of $c(U)$ on $\text{Sch}(U)$, the two classes of examples U and U' in Example 16 show that $c(U)$ is not lower bounded by a function of $\text{Sch}(U)$ with maximum or supremum value greater than 6. Whether $c(U)$ is upper bounded by a function of $\text{Sch}(U)$ is unknown, and this is listed as an open question in Sec. VIII.

A. Nonlocal cost of bipartite permutation operators

We have shown in Theorem 7 that every bipartite permutation operator can be implemented by three controlled unitaries, but such controlled unitaries may be hard to implement since they are on $d_A \times d_B$ space. A better measure of the nonlocal part of the gate cost (i.e., the nonlocal gate cost) is in terms of the bipartite elementary gates of Definition 13. Two results that depend on d_A and d_B are given in Proposition 15(iii)(iv), though special classes of the bipartite elementary gates are used therein. In this subsection we study the nonlocal gate cost as a function of the Schmidt rank or dimension of the bipartite permutation unitary. The obtained upper bounds could be much less than those in Proposition 15(iii)(iv) for some classes of permutation unitaries. The result can also be stated in terms of the entanglement cost under local operations and classical communications (LOCC). Hence it provides a significant class of examples that the entanglement cost of a bipartite unitary is upper bounded by a function of Schmidt rank independent of the dimensions. The only known result of this flavor is about Schmidt-rank-2 unitaries, which

are implementable using one ebit of entanglement under LOCC [15]. To study the nonlocal gate cost, we first present some definitions and preliminary lemmas.

Definition 17 A *partial permutation matrix* is a matrix with elements 0 or 1 only, with at most one nonzero element on each row and column. The input (respectively, output) space for such matrix is the complex Hilbert space that is the span of the computational basis states corresponding to the nonzero columns (respectively, rows) of the matrix.

Definition 18 The *partial-permutation rank* of a bipartite operator U , denoted $\text{ppr}(U)$, is the minimum number of terms q such that

$$U = \sum_{j=1}^q A_j \otimes B_j, \quad (33)$$

where A_j and B_j are partial permutation operators on \mathcal{H}_A and \mathcal{H}_B .

The above two definitions imply that if a bipartite operator has a partial-permutation rank, then its entries are non-negative integers. So the partial-permutation rank is not defined for a bipartite operator containing a negative or non-integer entry in its matrix. The partial-permutation rank of the bipartite permutation matrices will be studied in Lemma 21.

Lemma 19 Suppose U is a bipartite controlled unitary of the form $P_1 \otimes V_1 + P_2 \otimes V_2$, where P_1 and P_2 are orthogonal projectors on \mathcal{H}_A , and V_1 and V_2 are unitaries on \mathcal{H}_B . With the help of a one-qubit ancilla on each side, U can be implemented using two bipartite CNOT gates and some local unitary gates. The initial and final states of each ancilla qubit are the same.

Proof. Let a and b denote the qubit ancillas on the A and B side initialized in the state $|0\rangle_a$ and $|0\rangle_b$, respectively. The unitary U can be implemented using the ancillas and one CNOT gate with the following sequence of gates: a controlled gate on \mathcal{H}_{Aa} : $V_{Aa} = P_1 \otimes I_a + P_2 \otimes X_a$, where $X_a = |0\rangle\langle 1|_a + |1\rangle\langle 0|_a$ (similar below with subscripts changed), and a CNOT gate on \mathcal{H}_{ab} : $\text{CNOT}_{ab} = |0\rangle\langle 0| \otimes I_b + |1\rangle\langle 1| \otimes X_b$, and a controlled gate on \mathcal{H}_{bB} : $W_{bB} = |0\rangle\langle 0| \otimes V_1 + |1\rangle\langle 1| \otimes V_2$, and then CNOT_{ab} again to erase the state on b to $|0\rangle_b$, and the V_{Aa} again to erase the state on a to $|0\rangle_a$. This implements U without changing the states of a and b . \square

Lemma 20 Suppose U is a bipartite permutation unitary of partial-permutation rank q . Then the following statements hold:

- (i) with the help of a one-qubit ancilla on one party and a two-qubit ancilla on the other party, U can be implemented using at most $6q$ bipartite CNOT gates and some local permutation gates.
- (ii) with the help of a one-qubit ancilla on either party

and $3q$ ebits of entanglement, U can be implemented using LOCC.

Proof. (i) Consider the matrix representation of U in the computational basis of $\mathcal{H}_A \otimes \mathcal{H}_B$. Then U can be expanded as $U = \sum_{j=1}^q A_j \otimes B_j$, where A_j and B_j are partial permutation matrices. Each A_j or B_j has an input space and an output space as defined in Definition 17. Assume without loss of generality that the two-qubit ancilla is on the B side. Denote the ancilla qubit on the A side as a , and the two ancilla qubits on the B side as b and c . Let $\{|0\rangle, |1\rangle\}$ (with suitable subscripts a, b, c) be the computational basis of each ancilla qubit. Let the initial state of each of the ancilla qubits be $|0\rangle$.

Now let us visualize the computational basis of $\mathcal{H}_A \otimes \mathcal{H}_B$ as a rectangular table, with the rows labeling the computational basis states of \mathcal{H}_A , and columns for those of \mathcal{H}_B . The input spaces of $A_j \otimes B_j$ ($j = 1, \dots, q$) correspond to small (disconnected) rectangles in such a table. In other words, the computational basis states in the input space of $A_j \otimes B_j$ take all intersections of some rows and some columns of the table. In the following we abbreviate the word “disconnected” since it turns out that the rows and columns in such a “small rectangle” need not be consecutive in our argument.

The whole table of size $d_A \times d_B$ is thus partitioned into q disjoint small rectangles. The output spaces of $A_j \otimes B_j$ also correspond to small rectangles in the table. Our goal is to move the small rectangles to their respective desired positions, while for each such rectangle, we also hope to do an internal permutation of elements according to the form of A_j and B_j . Such internal permutation of elements is the tensor product of two permutations on a subspace of \mathcal{H}_A and a subspace of \mathcal{H}_B , respectively. But given that there may be some overlap between the input rectangle for one j and the output rectangle for a different j , it is hard to do an in-place swap of the rectangles. We avoid this problem by making use of the ancilla qubit c , since it effectively supplies two copies of the whole table of size $d_A \times d_B$, corresponding to the states $|0\rangle_c$ and $|1\rangle_c$, respectively. The latter copy is called the *backup copy* below.

For each $j = 1, \dots, q$, we perform the following procedure which consists of 3 controlled-permutation gates. Denote by P the rectangle corresponding to the input space of $A_j \otimes B_j$ in the original copy of the table. We first do a controlled-permutation unitary controlled in the computational basis of \mathcal{H}_A to swap the elements in P into the place (denoted by M) in the backup copy of the table and in the target columns. Then perform a controlled-permutation unitary controlled in the computational basis of $\mathcal{H}_B \otimes \mathcal{H}_c$ to swap the block M into the desired rows in the backup copy (denote the target rectangle by Q). Now the part of M that is not in Q (denoted by $M \setminus Q$) is an all-zero block (for any input state of the form $|\psi\rangle_{AB} \otimes |0\rangle_c$), but it should have the original contents before these two gates were applied, as it is not the output position for the original P . Therefore, we lastly perform a controlled unitary controlled in

the computational basis of \mathcal{H}_A to swap the partial rectangle $M \setminus Q$ and its corresponding part in P . Note that if $M \setminus Q$ is an empty set, the last two unitary gates are actually the identity operation. After the 3 gates, the original probability amplitudes in $M \setminus Q$ are unchanged, but the probability amplitudes in P and Q are swapped. Since the original state had zero probability amplitude in Q in the backup copy (note this is still true for $j > 1$ according to our procedure here), the state after these 3 gates has zero probability amplitude in the rectangle P (in the original copy). The internal permutations required in each rectangular block can be accomplished in the first two of these three controlled-permutation gates.

After performing the (at most) $3q$ controlled-permutation gates, we do a local X_c gate on particle c to swap the states $|0\rangle_c$ and $|1\rangle_c$. Now the U is implemented and the ancilla qubits are back in their original state. The local gate X_c can be absorbed into the last one of those (at most) $3q$ gates, which is a bipartite controlled-permutation gate controlled in the computational basis of \mathcal{H}_A . Thus U is implemented using at most $3q$ controlled-permutation unitaries, with the help of the ancilla qubit c . Lemma 19 implies that each of these controlled-permutation gates, which can be written in two terms, can be implemented using two bipartite CNOT gates and some local gates, the latter are local permutation gates in the current case. The two ancilla qubits used are a and b , and they can be recycled through these applications of Lemma 19, because they start and end in the $|0\rangle$ state in each application. Hence at most $6q$ bipartite CNOT gates and some local permutation gates can implement U with the help of the ancilla qubits a , b and c . This completes the proof of (i).

(ii) The proof is similar to (i), but note that each of the (at most) $3q$ bipartite controlled-permutation gate with two terms can be implemented using 1 ebit of entanglement and LOCC [31]. Thus we need at most $3q$ ebits in total and also need the ancilla qubit c , but do not need the ancillary qubits a and b . This completes the proof of (ii). \square

Next we relate the partial-permutation rank with the Schmidt rank.

Lemma 21 *Suppose the bipartite permutation unitary has partial-permutation rank q and Schmidt rank r . Then $q \leq \min\{d_A^2, d_B^2, d_A r, d_B r, 2^r\}$.*

Proof. Suppose U is the bipartite permutation unitary on the $d_A \times d_B$ system. The matrix U consists of d_A^2 blocks, each of which is a $d_B \times d_B$ partial permutation matrix. We denote them by B_{jk} , so that $U = \sum_{jk} |j\rangle\langle k| \otimes B_{jk}$. Since $|j\rangle\langle k|$ is also a partial permutation matrix, we have $q \leq d_A^2$ and by symmetry $q \leq d_B^2$. Since U is a permutation matrix, the nonzero blocks B_{jk} for fixed j are linearly independent. So the number of them is not greater than r . It holds for $j = 1, 2, \dots, d_A$. Thus the total number of nonzero B_{jk} is not greater than $d_A r$. Therefore $q \leq d_A r$. By symmetry of the A and B sides,

we have $q \leq d_{Br}$.

It remains to prove $q \leq 2^r$. Suppose $\{F_i\}_{i=1}^r$ is a set of r linearly independent blocks among B_{jk} . All other blocks that are not included in the set $\{F_i\}_{i=1}^r$ are linear combinations of the F_j . This last property does not change if we replace $\{F_i\}_{i=1}^r$ by $\{G_i\}_{i=1}^r$. Here each G_i is a linear combination of the F_j , and is of the standard form $G_i(t) = \delta_{it}$, $i = 1, 2, \dots, r$, where $G_i(t)$ is the t -th matrix element of G_i according to some fixed ordering of the matrix elements, and δ_{it} is the Kronecker delta function. Note that such orderings of matrix elements must exist but may not be the usual row-first ordering, as it depends on the operators F_i . Then any B_{jk} as a linear combination of G_i ($i = 1, 2, \dots, r$) must satisfy that the coefficient for G_i is either 0 or 1, since the resulting matrix is a $(0, 1)$ -matrix which implies that its first r elements (in the ordering above) must be either 0 or 1. Thus there are at most 2^r choices of the ordered set of coefficients, leading to at most 2^r distinct blocks B_{jk} . Denote the distinct B_{jk} as D_l , $l = 1, 2, \dots, m$. Then $m \leq 2^r$, and $U = \sum_{l=1}^m \left(\sum_{(j,k) \in S_l} |j\rangle\langle k| \right) \otimes D_l$, where $S_l = \{(j, k) : B_{jk} = D_l\}$. Since U is a permutation matrix, the operators $\sum_{(j,k) \in S_l} |j\rangle\langle k|$ are partial permutations, for any l . Hence $q \leq m \leq 2^r$. This completes the proof. \square

Lemmas 20 and 21 immediately imply

Theorem 22 *Suppose U is a bipartite permutation unitary of Schmidt rank r . With the help of a one-qubit ancilla on one party and a two-qubit ancilla on the other party, U can be implemented using at most $6 \min\{d_A^2, d_B^2, d_{Ar}, d_{Br}, 2^r\}$ bipartite CNOT gates and some local permutation gates. Alternatively, U can be implemented using LOCC and at most $3 \min\{d_A^2, d_B^2, d_{Ar}, d_{Br}, 2^r\}$ ebits of entanglement with the help of one ancillary qubit on either party.*

The theorem can also be stated for classical reversible circuits, by making minor changes such as replacing “qubit” with “bit”, and “local permutation gates” with “local reversible gates”.

In Lemma 21, the dimension-independent upper bound 2^r may still be improved. However, the following Example 24 provides evidence that at least for a class of bipartite permutation matrices, the partial-permutation rank grows fast with r (but is not known to be exponential). The operational meaning of the Schmidt rank r is that its logarithm is an upper bound of how many ebits of entanglement a bipartite unitary can create starting from a product state (possibly with local ancillas). Thus the separation between the partial-permutation rank and the Schmidt rank gives some indication about the separation of the entangling power and the entanglement cost under our protocol in the proof of Lemma 20.

Before presenting the example, we first define some versions of ranks for matrices. Let $\text{rank}(T)$ denote the usual rank of a matrix T .

Definition 23 (i). *For a matrix T with nonnegative elements, the nonnegative rank [32] $\text{rank}^+(T)$ is the minimum number of rank-1 matrices with nonnegative elements that sum to T .*

(ii). *For a binary matrix T (binary means the elements are 0 or 1), the binary rank [33] $\text{rank}_N(T)$ is the minimum number of rank-1 binary matrices that sum to T .*

(iii). *For a binary matrix T , the XOR rank [33] $\text{rank}_X(T)$ (also called modulo-2 rank) is the minimum number of rank-1 binary matrices such that their sum modulo 2 is T . It is also equal to the rank over the finite field \mathbb{F}_2 , or the number of linearly independent rows (or columns) under arithmetic operations in \mathbb{F}_2 .*

It is apparent that $\text{rank}(T) \leq \text{rank}^+(T) \leq \text{rank}_N(T)$ and $\text{rank}_X(T) \leq \text{rank}_N(T)$ hold for any binary matrix T , and according to [33], $\text{rank}_X(T)$ and $\text{rank}(T)$ are generally incomparable.

Example 24 Consider bipartite permutation unitaries of the form

$$U = \sum_{i=1}^M (|2i-1\rangle\langle 2i-1| + |2i\rangle\langle 2i|) \otimes (I_B - C_i) + (|2i-1\rangle\langle 2i| + |2i\rangle\langle 2i-1|) \otimes C_i, \quad (34)$$

where C_i are $d_B \times d_B$ diagonal partial permutation matrices, $i = 1, 2, \dots, M$. The diagonal part of U is $U_{\text{diag}} = \sum_{i=1}^M (|2i-1\rangle\langle 2i-1| + |2i\rangle\langle 2i|) \otimes (I_B - C_i)$. It is a partial permutation matrix, but its elements are all diagonal so the implementation is trivial, thus the implementation cost of U in terms of the protocol in the proof of Lemma 20 is determined by the off-diagonal part of U , which is denoted $U_{\text{od}} := \sum_{i=1}^M (|2i-1\rangle\langle 2i| + |2i\rangle\langle 2i-1|) \otimes C_i$. Hence $\text{ppr}(U_{\text{od}})$ is proportional to the non-local implementation cost under our protocol. The diagonal elements of the matrices C_i can be rearranged into a matrix T of size $M \times d_B$ with elements $T_{jk} = \langle k|_B C_j |k\rangle_B$. It is known that $\text{rank}^+(T) \leq \text{rank}_N(T)$. And $\text{rank}_N(T) = \text{ppr}(U_{\text{od}})$, since the minimum-term expansion of U_{od} of the form (33) must involve local operators on \mathcal{H}_A which are the tensor product of a diagonal partial permutation operator on an M -dimensional space and the operator $|1\rangle\langle 2| + |2\rangle\langle 1|$ on a two-dimensional space, and the partial permutation operators on B are diagonal. These two types of diagonal partial permutation operators mentioned above correspond to the column and row vectors in an expansion of T in terms of direct products of binary column and row vectors. Therefore $\text{rank}^+(T) \leq \text{rank}_N(T) = \text{ppr}(U_{\text{od}})$. On the other hand, $\text{rank}(T) \geq \text{Sch}(U_{\text{od}})$, since the expansion of T using $\text{rank}(T)$ terms which are the direct products of a column vector and a row vector, corresponds to an expansion of U_{od} in terms of tensor-product operators. Therefore, any separation between $\text{rank}(T)$ and $\text{rank}^+(T)$ provides a lower bound for the separation between $\text{Sch}(U_{\text{od}})$ and $\text{ppr}(U_{\text{od}})$. By the way, for this U we have $|\text{Sch}(U_{\text{od}}) - \text{Sch}(U)| \leq 1$, since the diagonal blocks

are $I_B - C_i$, and the Schmidt rank is equal to the number of linear independent $d_B \times d_B$ blocks in the matrix U . The operational meaning of $\text{Sch}(U)$ is mentioned before the example.

From [34], there is a class of $(0,1)$ -matrices T such that the separation between $\text{rank}_\epsilon(T)$ and of $\text{rank}_\epsilon^+(T)$ is at least quasipolynomial (but not known to be exponential), more precisely, $\log \text{rank}_\epsilon^+(T) \geq \Omega(\log^2 \text{rank}_\epsilon(T))$ for such T . The subscript ϵ means the T could be replaced by a matrix that approximates T to accuracy ϵ for evaluation of the rank, which makes sense in terms of physical implementation. Therefore the separation between $\text{Sch}_\epsilon(U_{od})$ and $\text{ppr}_\epsilon(U_{od})$ is at least quasipolynomial for a certain class of permutation matrix U , where the subscript ϵ has the same meaning as above.

It is interesting to note that the problem of the separation of the rank and nonnegative rank is related to the log-rank conjecture [35] in communication complexity theory (as remarked in [34]). It is curious that the nonlocal cost for implementing bipartite permutation unitaries (or reversible circuits) under our protocol is related to the communication complexity theory in this unexpected way. \square

The lower bound mentioned above is not polynomial. So our protocol in the proof of Lemma 20 is not efficient for some subclass of bipartite permutation unitaries represented by Eq. (34). There is an alternative method of implementing these unitaries, illustrated in Example 25 below, by noting that the local gate $|2i-1\rangle\langle 2i| + |2i\rangle\langle 2i-1|$ applied twice is the projector on the two-dimensional subspace spanned by $|2i-1\rangle$ and $|2i\rangle$. Rather than expanding U_{od} using the form (33), we can write U as the product of some controlled-permutation unitaries controlled in the computational basis of \mathcal{H}_B , where the controlled operators on \mathcal{H}_A are either I_A , or a permutation unitary which is the direct sum of an identity operator on a two-dimensional subspace, with another operator which is the tensor product of the identity operator on an $(M-k)$ -dimensional space and the operator $|1\rangle\langle 2| + |2\rangle\langle 1|$ on a two-dimensional space. Each such controlled-permutation gate can be implemented using two bipartite CNOT gates and some local unitaries with the method in Lemma 19. The nontrivial part of these controlled-permutation unitaries could overlap with each other. It corresponds to that in the expansion of T in terms of binary vectors, the operator XOR is to be used instead of the usual “+” operator. That is, $T = \oplus_j (u_j \otimes v_j)$, where u_j and v_j are binary column and row vectors, respectively, and \oplus represents element-wise XOR operation (modulo-2 addition) of two or more matrices. Therefore, the relation between the XOR-rank of binary matrices and the rank of these matrices is relevant for the separation of the implementation cost and the Schmidt rank of U . In general, there is no definite inequality relation between the XOR-rank and the rank of a binary matrix [33]. Thus there may be some cases where this modified protocol is quite efficient, but in the other cases it is not too bad either, since $\text{rank}_X(T) \leq \text{rank}_N(T)$ holds for any binary matrix

T , meaning that it is better than the original protocol.

Example 25 Let U be of the form in Eq. (34), where $d_A = 6$, $M = d_B = 3$, and $C_1 = \text{diag}(1, 1, 0)$, $C_2 = \text{diag}(1, 0, 1)$, $C_3 = \text{diag}(0, 1, 1)$. Let $V = (X \oplus X \oplus I_2) \otimes \text{diag}(1, 1, 0) + I_6 \otimes \text{diag}(0, 0, 1)$, where $X = |1\rangle\langle 2| + |2\rangle\langle 1|$, and I_n is the $n \times n$ identity matrix. Let $W = (I_2 \oplus X \oplus X) \otimes \text{diag}(0, 1, 1) + I_6 \otimes \text{diag}(1, 0, 0)$. Then $U = VW$. Each of V and W can be implemented by two bipartite CNOT gates (ignoring local unitary gates; same as below). Hence U can be implemented by four bipartite CNOT gates. In comparison, the protocol in the proof of Lemma 20, enhanced by doing nontrivial operations for the off-diagonal part U_{od} only, requires 6 bipartite CNOT gates (the partial permutation rank of U_{od} is $q = 3$, and a reduction by a factor of 3 applies because the partial permutations are in-place). The corresponding T matrix is

$$T = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}. \quad (35)$$

It has rank 3, and XOR rank 2. This is the simplest example of a binary matrix that has its XOR rank less than the rank.

VII. THE CASE WITH ANCILLAS

The use of ancillas of constant size has been seen in the previous section. In this section, we show that the use of ancillas of variable size (sometimes required to be initialized in fixed states) can be useful for reducing the controlled-gate cost $c(U)$ or the number of CNOT gates needed, but sometimes at the cost of modifying the U (e.g., the tensor product $U \otimes I_G$ instead of U itself is used in Proposition 27 below).

Proposition 26 *Any bipartite unitary U on $d_A \times d_B$ system can be implemented by $4\lceil \log_2 \min\{d_A, d_B\} \rceil$ bipartite CNOT gates and some local unitaries, with the help of $\lceil \log_2 \min\{d_A, d_B\} \rceil$ ancilla qubits.*

Proof. The circuit for implementing U is as follows: send the state of one system (which is embedded in an integer number of qubits) to the other party using $2\lceil \log_2 \min\{d_A, d_B\} \rceil$ CNOT gates, then perform the U locally, and finally send the state of the said system back using the inverse of the first part of the circuit. The first part of the circuit is a tensor product of many subcircuits each sending one qubit. Each such subcircuit is exactly the one-bit teleportation circuit in [36], and requires one ancilla qubit which is initially in a fixed state and finally contains the one-qubit state being transferred. The number of subcircuits in the first part of the whole circuit is $\lceil \log_2 \min\{d_A, d_B\} \rceil$, and since final transfer back to the first system reuses the original qubits, no extra ancillas

are needed, therefore the total number of ancilla qubits needed is $\lceil \log_2 \min\{d_A, d_B\} \rceil$. \square

Some special classes of bipartite unitaries can be implemented with small amounts of entanglement and classical communication [31], which can also be expressed in terms of CNOT gates. It should be noted that the upper bound $4\lceil \log_2 \min\{d_A, d_B\} \rceil$ is not optimal for all dimensions: as mentioned previously, in the case of $d_A = d_B = 2$, only 3 CNOT gates together with local unitaries are needed, without using ancillas. We do not know whether this upper bound is optimal for general unitaries in other dimensions, and this is listed as an open question in the next section.

Somewhat surprisingly, Example 16 in Sec. VI shows the following:

Proposition 27 *For any bipartite unitary U , $c(U) \geq c(U \otimes I_G)$, where G is one qubit on the A side. There are examples of U satisfying $c(U) > c(U \otimes I_G)$.*

Proof. The inequality is from observing that any decomposition of U using controlled unitaries can be extended to a decomposition of $U \otimes I_G$ with the same number of controlled unitaries. If $c(U) = c(U \otimes I_G)$ always holds, we may repeatedly use it by adding one qubit on the A side at a time, and get $c(U) = c(U \otimes I_{A'})$, where A' has an integer number of qubits and is of size at least as big as A . Then the method in Example 16 implies that $c(U \otimes I_{A'}) \leq 6$, thus $c(U) \leq 6$, but this is generally impossible for a generic U simply by parameter counting, see Sec. III. Therefore $c(U) = c(U \otimes I_G)$ does not always hold. \square

VIII. CONCLUSIONS

We have proposed the sandwich and generalized sandwich forms for the decomposition of bipartite and multipartite unitary operators, respectively. In particular, we have shown that any bipartite unitary on $\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}$ has a $(4d_A - 5)$ -sandwich form, and any n -partite unitary on $\mathbb{C}^{d_1} \otimes \dots \otimes \mathbb{C}^{d_n}$ has a generalized $[2 \prod_{j=1}^{n-1} (2d_j - 2) - 1]$ -sandwich form. The numbers can be further reduced in some special cases. In particular, three controlled unitaries can implement a bipartite complex permutation operator. This last result can be applied to classical reversible circuits. We mentioned some connections between our results and the results in the literature. As an application of the types of decompositions above, we discussed how to express a bipartite unitary as the product of a simple type of bipartite gates and some local unitaries. We also discussed the relationship between the Schmidt rank of the unitary (bipartite permutation unitary in particular) and the complexity of the decomposition, and also discussed the use of local ancillas. To conclude this paper we present a few open questions by requiring that the gates are exactly implemented, and no ancillary space or system is allowed unless stated otherwise.

1. Let $s(U)$ be the smallest number of controlled unitary gates required in a decomposition of U of the sandwich form. Then $s(U) \geq c(U)$. Do we have $s(U) = c(U)$? We suspect that this does not hold for some U . But does the similar equality $\max_{U \in \mathcal{T}_{ab}} s(U) = \max_{U \in \mathcal{T}_{ab}} c(U)$ hold, where \mathcal{T}_{ab} is the set of all bipartite unitaries U on an $a \times b$ dimensional space?
2. Can we obtain some form of decomposition of bipartite unitaries in terms of controlled unitaries, by taking a hint from the decomposition of single-party unitary matrices in [37, Corollary 1]? In particular, can we replace $4d_A - 5$ by $2d_A - 1$ in (3)?
3. Let U be a bipartite unitary on the $d_A \times d_B$ system. Do the following equations hold?

$$\begin{aligned} c(U) &= c(U + |d_A + 1\rangle\langle d_A + 1| \otimes I_B), \\ c_s(U) &= c_s(U + |d_A + 1\rangle\langle d_A + 1| \otimes I_B), \\ c_e(U) &= c_e(U + |d_A + 1\rangle\langle d_A + 1| \otimes I_B). \end{aligned} \quad (36)$$

4. It is obvious that the following inequalities hold:

$$\begin{aligned} c(U^{\otimes n}) &\leq c(U), \\ c_s(U^{\otimes n}) &\leq n \cdot c_s(U), \\ c_e(U^{\otimes n}) &\leq n \cdot c_e(U). \end{aligned} \quad (37)$$

Here the bipartite unitary $U^{\otimes n} = U_{A_1 B_1} \otimes \dots \otimes U_{A_n B_n}$ acts on the space $\mathcal{H}_A \otimes \mathcal{H}_B$ where $\mathcal{H}_A = \bigotimes_{i=1}^n \mathcal{H}_{A_i}$ and $\mathcal{H}_B = \bigotimes_{i=1}^n \mathcal{H}_{B_i}$. But do the equalities always hold in the three inequalities above?

5. As discussed in Sec. V, a bipartite permutation unitary can be implemented using a certain number of standard gates of the CNOT type and some local transposition gates. What is the minimum number of these gates needed to implement any bipartite permutation unitary on $d_A \times d_B$ space? And since the local gates can be regarded as easy to implement, we can also ask the following: What is the minimum number of the first type of gates needed?
6. For d_A or d_B greater than 2, can an upper bound better than $4\lceil \log_2 \min\{d_A, d_B\} \rceil$ be found for the number of CNOT gates (or the bipartite elementary gates defined in Definition 13) needed to decompose a bipartite unitary with the help of local ancillas and local unitaries?
7. Is there a dimension-independent upper bound of $c(U)$ in terms of the Schmidt rank r of a general bipartite unitary U ? It is known that $c(U) = 1$ when $r = 2$ or 3 [15, 17]. See also the discussions in Sec. VI, and Theorem 22 (which is for the permutation unitaries only and requires ancillas of fixed size).

8. For given integers m, n satisfying $m > n \geq 2$, is there a dimension-independent upper bound (as a function of m, n only) of the number of Schmidt-rank- n bipartite unitaries needed to decompose any Schmidt-rank- m bipartite unitary on the same space? What about restricting the target unitary to be a controlled unitary in this question?

Acknowledgments

We thank Joseph Fitzsimons and Kae Nemoto for helpful discussions. We also thank Scott Cohen for careful

reading of an early version of the paper and pointing out a couple of issues in the presentation. This material is based on research funded in part by the Singapore National Research Foundation under NRF Grant No. NRF-NRFF2013-01, and in part by NICT-A (Japan).

-
- [1] Michael Reck, Anton Zeilinger, Herbert J. Bernstein, and Philip Bertani. Experimental realization of any discrete unitary operator. *Phys. Rev. Lett.*, 73:58–61, Jul 1994.
- [2] J.-L. Brylinski and R. Brylinski. *Mathematics of Quantum Computation*, edited by R. Brylinski and G. Chen, CRC Press, 2002.
- [3] Stephen S. Bullock, Dianne P. O’Leary, and Gavin K. Brennen. Asymptotically optimal quantum circuits for d-level systems. *Phys. Rev. Lett.*, 94:230502, Jun 2005.
- [4] Yu Chen, C. Neill, P. Roushan, N. Leung, M. Fang, R. Barends, J. Kelly, B. Campbell, Z. Chen, B. Chiaro, A. Dunsworth, E. Jeffrey, A. Megrant, J. Y. Mutus, P. J. J. O’Malley, C. M. Quintana, D. Sank, A. Vainsencher, J. Wenner, T. C. White, Michael R. Geller, A. N. Cleland, and John M. Martinis. Qubit architecture with high coherence and fast tunable coupling. *Phys. Rev. Lett.*, 113:220502, Nov 2014.
- [5] Throughout the paper a local unitary is in general not a single-qubit unitary; “local” is with respect to the system partition.
- [6] G. K. Brennen, S. S. Bullock, and D. P. O’Leary. Efficient circuits for exact-universal computation with qudits. *Quantum Info. Comput.*, 6(4):436–454, July 2006.
- [7] Yao-Min Di and Hai-Rui Wei. Synthesis of multivalued quantum logic circuits by elementary gates. *Phys. Rev. A*, 87:012325, Jan 2013.
- [8] V.V. Shende, S.S. Bullock, and I.L. Markov. Synthesis of quantum-logic circuits. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 25(6):1000–1010, June 2006.
- [9] Yaoyun Shi. Quantum and classical tradeoffs. *Theoretical Computer Science*, 344(23):335 – 345, 2005.
- [10] V.V. Shende, A.K. Prasad, I.L. Markov, and J.P. Hayes. Synthesis of reversible logic circuits. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 22(6):710–722, June 2003.
- [11] Mehdi Saeedi and Igor L. Markov. Synthesis and optimization of reversible circuits - a survey. *ACM Computing Surveys*, 45, 2, Article 21 (34 pages), 2013.
- [12] Farrokh Vatan and Colin Williams. Optimal quantum circuits for general two-qubit gates. *Phys. Rev. A*, 69:032315, Mar 2004.
- [13] Michael A. Nielsen, Christopher M. Dawson, Jennifer L. Dodd, Alexei Gilchrist, Duncan Mortimer, Tobias J. Osborne, Michael J. Bremner, Aram W. Harrow, and Andrew Hines. Quantum dynamics as a physical resource. *Phys. Rev. A*, 67:052301, May 2003.
- [14] Jon Tyson. Operator-Schmidt decompositions and the Fourier transform, with applications to the operator-Schmidt numbers of unitaries. *J. Phys. A: Math. Gen.*, 36:10101, 2003.
- [15] Scott M. Cohen and Li Yu. All unitaries having operator Schmidt rank 2 are controlled unitaries. *Phys. Rev. A*, 87:022329, Feb 2013.
- [16] Lin Chen and Li Yu. Nonlocal and controlled unitary operators of Schmidt rank three. *Phys. Rev. A*, 89:062326, Jun 2014.
- [17] Lin Chen and Li Yu. On the Schmidt-rank-three bipartite and multipartite unitary operator. *Annals of Physics*, 351:682–703, 2014.
- [18] Juan Carlos Garcia-Escartin and Pedro Chamorro-Posada. A SWAP gate for qudits. *Quantum Information Processing*, 12(12):3625–3631, 2013.
- [19] Navin Khaneja, Roger Brockett, and Steffen J. Glaser. Time optimal control in spin systems. *Phys. Rev. A*, 63:032308, Feb 2001.
- [20] B. Kraus and J. I. Cirac. Optimal creation of entanglement using a two-qubit gate. *Phys. Rev. A*, 63:062309, May 2001.
- [21] C.C. Paige and M. Wei. History and generality of the CS decomposition. *Linear Algebra and its Applications*, 208–209:303–326, 1994.
- [22] Stephen Brierley and Stefan Weigert. Constructing mutually unbiased bases in dimension six. *Phys. Rev. A*, 79:052316, May 2009.
- [23] Andrew S. Maxwell and Stephen Brierley. On properties of Karlsson Hadamards and sets of mutually unbiased bases in dimension six. *Linear Algebra and its Applications*, 466(0):296 – 306, 2015.
- [24] Jonathan Welch, Alex Bocharov, and Krysta M. Svore. Efficient Approximation of Diagonal Unitaries over the Clifford+T Basis. <http://arxiv.org/abs/1412.5608>, December 2014.
- [25] Hall’s marriage theorem. http://en.wikipedia.org/wiki/Hall's_marriage_theorem.
- [26] Alexis De Vos. *Reversible Computing: Fundamentals, Quantum Computing, and Applications*. Wiley-VCH, 2010.

- [27] Cheng Peng, G.V. Bochmann, and T.J. Hall. Quick Birkhoff-von Neumann Decomposition Algorithm for Agile All-Photonic Network Cores. In *Communications, 2006. ICC '06. IEEE International Conference on*, volume 6, pages 2593–2598, June 2006.
- [28] Alexis De Vos. Reversible computer hardware. *Electronic Notes in Theoretical Computer Science*, 253(6):17 – 22, 2010. Proceedings of the Workshop on Reversible Computation (RC 2009).
- [29] T. C. Ralph, K. J. Resch, and A. Gilchrist. Efficient Toffoli gates using qudits. *Phys. Rev. A*, 75:022313, Feb 2007.
- [30] Benjamin P. Lanyon, Marco Barbieri, Marcelo P. Almeida, Thomas Jennewein, Timothy C. Ralph, Kevin J. Resch, Geoff J. Pryde, Jeremy L. O’Brien, Alexei Gilchrist, and Andrew G. White. Simplifying quantum logic using higher-dimensional Hilbert spaces. *Nat Phys*, 5(2):134–140, 02 2009.
- [31] Li Yu, Robert B. Griffiths, and Scott M. Cohen. Efficient implementation of bipartite nonlocal unitary gates using prior entanglement and classical communication. *Phys. Rev. A*, 81:062315, Jun 2010.
- [32] Nonnegative rank (linear algebra). [http://en.wikipedia.org/wiki/Nonnegative_rank_\(linear_algebra\)](http://en.wikipedia.org/wiki/Nonnegative_rank_(linear_algebra)).
- [33] Pauli Miettinen. Binary matrix factorisations (tutorial @ ECML PKDD 2012). http://people.mpi-inf.mpg.de/~pmiettinen/bmf_tutorial/tutorial.html.
- [34] Gillat Kol, Shay Moran, Amir Shpilka, and Amir Yehudayoff. Approximate nonnegative rank is equivalent to the smooth rectangle bound. In Javier Esparza, Pierre Fraigniaud, Thore Husfeldt, and Elias Koutsoupias, editors, *Automata, Languages, and Programming*, volume 8572 of *Lecture Notes in Computer Science*, pages 701–712. Springer Berlin Heidelberg, 2014.
- [35] László Lovász and Michael E. Saks. Lattices, Möbius functions and communication complexity. In *FOCS*, pages 81–90. IEEE Computer Society, 1988.
- [36] Xinlan Zhou, Debbie W. Leung, and Isaac L. Chuang. Methodology for quantum logic gate construction. *Phys. Rev. A*, 62:052316, Oct 2000.
- [37] Martin Idel and Michael M. Wolf. Sinkhorn normal form for unitary matrices. *Linear Algebra and its Applications*, 471(0):76 – 84, 2015.